A Report to the Library of Congress

WEB PRESERVATION PROJECT
INTERIM REPORT

William Y. Arms
Cornell University

January 15, 2001

The work described in this report was carried out during summer and fall of 2000 by Roger Adkins, Cassy Ammen and Allene Hayes from the Library Congress, with William Arms of Cornell University as consultant. Melissa Levine provided expertise on copyright and other legal issues. The team met regularly with an advisory group of Barbara Tillett, Jane Mandelbaum and Diane Kresh. This report was written by William Arms who takes responsibility for the contents. It makes extensive use of notes, comments and observations provided by the other members of the project.

Contents

1. Summary of Recommendations

This report to the Library of Congress has two functions. The first is to describe the pilot phase of the Web Preservation Project, an experiment in collecting materials that are available with open access on the web. The second is to discuss and make recommendations about collecting and preserving such materials for the long term.

1.1 General recommendations

This work has successfully demonstrated the processes by which the Library can select, collect, organize and preserve open access materials from the web by downloading copies over the Internet.

- There are no serious impediments to the Library undertaking a broad program of collecting web sites, as part of its mission to collect and preserve the cultural and intellectual artifacts of today for the benefit of future generations. However, this is a substantial undertaking, which is not worth beginning without a dedicated team of librarians and technical staff, about ten people initially.

- Collecting, organizing, preserving and providing access to web materials are interrelated. Several of these activities overlap with other developments at the Library and can share planning and expertise.

- To share the total effort, the Library can benefit from partnerships with other libraries and archives.

- Because of the volume of web materials that might be collected, most processes will be automatic. Skilled librarians are needed to establish and monitor the procedures and to give special attention to a small number of particularly important web sites.

- Collecting, preserving and providing access to huge volumes of online information is technically challenging. Some parts of the process can be subcontracted to specialists, but the Library needs to establish the expertise to manage the overall process.

1.2 Selection and collection

There are two approaches to selection and collection. Bulk collecting is largely automated. It is an economical way to collect very large amounts of material. Selective collecting is managed by librarians who select, index and organize materials for collection and preservation.

- A comprehensive strategy for the Library of Congress will combine bulk collection and selective collection. Technically, both are sensible approaches. For future scholars, selective collection offers the benefits of depth whereas bulk collection provides breadth of coverage at much lower cost.

- The cost and complexity of collecting a web site is influenced by a number of selection decisions, for example the frequency of collecting each site and whether to collect the entire site or only certain data types such as text and images.

1.3 Use of the collections for scholarship and research

The study has examined two ways in which the collections might be used in the future for scholarship and research. Both begin with the source file that is downloaded over the web. The more basic method is to use computer programs to analyze the unmodified source files. The alternative is to create a web site that allows a researcher to view the downloaded web sites as they were on the day when they were captured. This requires editing the source files to create a version for access.

- These two types of analysis -- by computer using the source files and by viewing an access version of the web sites -- serve different types of research. The Library will need to provide both.

- Editing files to create the access version will have to be carried out automatically. Because of the ever-changing technology of the web, the Library cannot guarantee that the access version will exactly replicate the experience of viewing the original site.

- These materials are subject to copyright. Therefore, the Library will have to establish policies and procedures for who has access to the collections under what conditions.

1.4 Information discovery

Since very large numbers of web sites will be collected and preserved, some form of catalog, index or finding aid is required, but here is no body of experience to know how researchers will use a large archive of web materials. For this reason, any plans for cataloguing and indexing are highly tentative.

- The Library is encouraged to experiment with various approaches to indexing and cataloguing web sites, including automated indexing, Dublin Core and MARC cataloguing. In particular, the Library needs to study the effectiveness of automated indexing as a low-cost alternative to cataloguing.

- The Library will probably not be able to afford full original catalog records for all web sites that are collected.

- Approaches to indexing and cataloguing must recognize that web sites change continually, so that catalog records will need continuous maintenance.

The pilot study successfully created MARC records for web sites and also provided a list of sites by URL, a list of sites by title and groupings by subject headings. It did not experiment with any form of automatic indexing.

1.5 The legal situation

While the collection of these digital materials for the benefit of future scholarship is consistent with the Library's duties, the Copyright Act does not reflect the current situation adequately.

- The Library should work with the Copyright Office to identify the necessary changes and to have them reflected by appropriate regulations and legislation.

- For collecting open access information from the web, the Library needs the clearly stated right to download materials without explicit permission from the copyright owner, and the right to have materials collected by other organizations working as the Library's agents.

- In order for the Library to provide scholars and other patrons with access to open access materials collected from the web, it needs the right to do so, under appropriate controls.

1.6 Long-term preservation

It is difficult to guess how future researchers will use preserved collections of web sites, yet strategies for preservation are interwoven with assumptions about why the collections are being preserved. The following recommendations apply to all digital materials that the Library holds in its collections.

- The Library should keep the original source material as collected and also any altered versions that have been edited for access versions or for preservation, complete with provenance metadata.

- Periodically, all versions need to be refreshed by copying from old media to new.

- On a regular basis, the staff will need to review the formats, protocols, program code, etc., and, if they are becoming obsolete, migrate to the nearest modern equivalent.

- The Library will need to maintain a library of technical specifications, standards and software utilities that describe the technology used by web sites. This will be a resource for future scholars in puzzling out file types and formats that have become obsolete.

Even with extensive migration, there is no guaranteed way to preserve the experience of using digital materials.

1.7 The development of a production system

The report contains a preliminary analysis of the resources needed to begin selective collection of 30,000 web sites. (This number is one percent of OCLC's estimate of the total number of web sites.)

- This is a substantial computing system that should be developed following modern software engineering practices, beginning with a full requirements analysis.

- The staffing requirements will be affected crucially by policy decision about selection, type of use by scholars, level of cataloguing or indexing, and preservation strategy.

- The components of this system need to be integrated with those for other digital materials that the Library has to manage.

2. Open Access Materials on the Web

2.1 Objective

Preserving open access materials from the web falls within the Library of Congress's mission to collect and preserve the cultural and intellectual artifacts of today for the benefit of future generations. An ever-increasing amount of primary source materials are being created in digital formats and distributed on the web, and do not exist in any other form. Future scholars will need them to understand the cultural, economic, political and scientific activities of today and, in particular, the changes that have been stimulated by computing and the Internet.

Two recent reports by the National Research Council have emphasized the important role of the Library of Congress in collecting and preserving these digital materials. *A Digital Strategy for the Library of Congress* praised the Library's plans to address the challenges and urged the Library to move ahead rapidly [1]. *The Digital Dilemma*, in its excellent chapter on "Public Access to the Intellectual, Cultural, and Social Record," noted that the current system of copyright deposit needs modification to enable such preservation [2].

2.2 Strategies for collecting open access web materials

The materials on the web can be divided into two categories, those that are provided with open access and those where there are access restrictions. This study looks at only the first category, <u>open access materials with no access restrictions</u>. The Library of Congress has other initiatives that are exploring strategies for collecting materials with restricted access.

The materials that are available on the Internet with open access are an important category of digital materials. These fragile materials are invaluable, yet quickly transient and disappearing records of our time. Of course, not everything is worthy of retention nor is an entirely comprehensive effort necessary or even useful to future researchers. Yet, the numbers of potentially valuable items is enormous and growing. In February 2000, OCLC estimated that were 2.9 million public web sites [3]. In July 2000, the Google search engine first reported that it indexed more than one billion web pages. A significant, but unknown proportion of these pages change every month.

Approaches for collecting open access materials from the web can be divided into two categories, depending on the degree of automation.

- <u>Bulk collecting</u> is largely automated. It is an economic way to collect very large amounts of material.

- <u>Selective collecting</u> is managed by librarians who select, index and organize materials for collection and preservation.

Both approaches rely heavily on automatic processes to download, store and preserve materials from the Internet. As a practical matter, automation is the only way to collect and manage the collections; these tasks are impossible if everything is carried out manually. The first phase of the Web Preservation Project has studied selective collecting, but, as discussed later, there are strong reasons, both scholarly and economic, for considering bulk and selective collecting as two facets of a single strategy.

2.3 The Minerva pilot

The purpose of the pilot phase of the Web Preservation Project was to gain insights into the practical issues involved in collecting and organizing selected web sites. To achieve this, a small number of open access web sites were collected, catalog records created and a demonstration web site set up. This pilot was given the name "Minerva"[1].

The main activities in the pilot phase were as follows:

- A small number of web sites were nominated by selection officers at the Library. Three sites were chosen for close study:

  http://www.whitehouse.gov/
  http://www.algore2000.com/
  http://www.georgewbush.com/

- Snapshots of these sites were downloaded using the HTTrack mirroring program. The snapshots were inspected for errors, anomalies, etc.

- Catalog records were created using OCLC's CORC software and loaded into Library of Congress's ILS system.

- A trial web site was developed to evaluate user access. This allowed users to view each version of each website as it was downloaded. The site also allowed users to search the ILS, or lists of URLs, titles and subject headings, with links to the collection of web sites.

- Discussions were held with the Copyright Office on legal issues.

In parallel to the pilot phase of the Web Preservation Project, the Library of Congress requested the Internet Archive to conduct a pilot test by capturing a set of web sites related to the elections in fall 2000. This test will capture some 150 to 200 web sites on a daily basis over the period leading up to the Inauguration 2001. In addition to capturing the materials, the pilot is creating a simple index and a time-dependent viewer. This experiment is due to end in March 2001.

[1] Minerva is a light hearted acronym for "Mapping the Internet: the Electronic Resource Virtual Archive."

2.4 Web preservation by other libraries and archives

There are several important web preservation projects in other libraries and archives.

- The National Library of Australia has been one of the pioneers of selective collection in its excellent Pandora program [4]. Since being established in June 1996, Pandora has established an archive of selected Australian online publications, including web sites, developed a national approach to the long-term preservation of these publications and provided support for indexing and abstracting agencies. The design of the Minerva web site was modeled on Pandora.

- The Internet Archive is a not-for-profit organization in San Francisco that has been collecting the open access HTML pages on the web since 1996, approximately monthly [5]. As of March 2000, its holdings were 11 terabytes on tape, 2.5 terabytes on disk. A commercial company, Alexa, carries out the data gathering. Alexa donates its data to the archive when it is six months old. The Internet Archive is currently setting up all the data on disk for researchers.

- Another pioneer in bulk collecting web materials is the National Library of Sweden, which has been collecting all web materials with Swedish content for several years [6]. The program, known as Kulturarw3, also began in 1996. Its aim is to test methods of collecting, preserving and providing access to Swedish electronic documents, which are accessible online in such a way that they can be regarded as published. By August 2000, it had completed seven downloads of the Swedish web, some 65 million items about half of which are text documents.

3. Selection and Collection

3.1 Capture of web sites

The unit of collection and preservation is a web site, typically consisting of many files. The basic technique to is to download a snapshot of each site, at predetermined intervals, using a mirroring program that copies each file from the web server to a computer at a library or archive. Thus, the archive will collect and preserve a sequence of snapshots for each site.

Each snapshot consists of a set of files with associated metadata, such as the date on which the snapshot was taken. The snapshot may contain all the files at the site or a subset. (For example, the snapshot might include text files only.) In all web collection, the actual downloading and storing of the files are carried out automatically.

3.2 Selection decisions

Several selection decisions must be made.

Which sites should be collected?

With selective collecting, the sites are determined by selection librarians or other experts. With bulk collecting, all sites that satisfy standard criteria are collected. (For example, the bulk selection decision might be made to collect all sites in the .gov domain. Alternatively, the Library might decide not to collect the .gov domain, but rely on a partnership with the National Archives and Records Administration.)

How often should snapshots be made?

The appropriate frequency for taking snapshots is a selection decision, which depends on the site. (For example, a site for a special event may be collected daily during the event, but at less frequent intervals before and afterwards.)

Which content should be collected?

Intuitively, a snapshot should include all files that are part of a site, but sometimes it is reasonable to be selective. Most of the largest files are in special formats, such as audio and video files; much of the complexity lies in the files that contain executable computer programs. (For example, in building their indexes of the web, web search programs ignore all files except HTML.)

3.3 Selection of web sites in the Minerva pilot

At the beginning of the Web Preservation Project, a group of recommending officers at the Library met to suggest possible web sites to select for the pilot phase. The guidelines used by the National Library of Australia were discussed, but it was decided to rely on the expertise of the librarians to nominate possible sites. A total of 35 sites were suggested. Of these, 29 were downloaded at least once. (They are listed in Appendix A.)

Three sites were studied in detail:

http://www.whitehouse.gov/
http://www.algore2000.com/
http://www.georgewbush.com/

3.4 Selection decisions by other web preservation projects

The Internet Archive and the Swedish Kulturarw3 project both have bulk collection policies for collecting open access web sites. Once a month, the Internet Archive collects every HTML page discovered by its web crawler, and associated images. It does not collect files in other formats, such as audio and video files, or program files. Kulturarw3 collects all web sites of interest to Sweden, including the entire .se domain. It collects the entire web sites.

The Australia Pandora has a selective collection policy. Librarians select web sites that have particular interest in Australia and decide the frequency of collection, separately for each site. Some sites are collected only once.

3.5 Collecting web sites

For Minerva, the program used to download the snapshots was HTTrack, a web crawler that is used to make mirrors of web sites. HTTrack is made openly available by its developers [7]. To run HTTrack, it is given a URL, usually the home page of a web site. The program makes a copy of the page and extracts all links to pages in the same web site. It then downloads those pages and continues until the entire web site has been copied. This process copies all files, including text, images, videos, executable programs, style sheets, metadata files, etc.

Collecting these sites illustrates several features that become challenges when collecting and preserving web sites:

Formats

The files are in a very wide variety of formats, for example text, images, audio or video. Many include segments of computer programs written in languages such as JavaScript or Java, or depend upon information stored in databases. Every year new formats and new versions of formats are introduced.

Boundaries

The boundaries of individual sites may be hard to define. The usual definition is to collect everything that has an address that is equal to or below the starting URL of the web site.

Errors

Many of the files contain errors and inconsistencies; they do not conform to the formats or contain hyperlinks (URLs) to files that do not exist. Many errors are identified by the mirroring program and recorded in a log file. As listed in Appendix A, every single site that was downloaded had at least one error.

System performance

HTTrack is a reasonably efficient mirroring program, yet, as shown in Appendix A, downloading many of the snapshots from a moderately powerful desktop computer with a good Internet connection took a long time, with a maximum approaching four hours.

Databases

Many websites draw content from a database. In some instances, every item in the database is accessible via its own URL and will be downloaded by mirroring. In other cases, notably search systems, the underlying database cannot be mirrored by programs such as HTTrack. This important category of material is beyond the scope of this report.

3.6 Selection recommendations for the Library of Congress

A comprehensive strategy for the Library of Congress should combine bulk collection and selective collection. Technically, both are sensible approaches. For future scholars selective collection offers the benefit of depth, whereas bulk collection provides breadth of coverage at much lower cost.

The cost difference is substantial. Comparing the staffing estimates in Section 8 with the experience of the Internet Archive suggests that selective collection as carried out by the Minerva pilot is about 100 times more expensive than bulk collection.

However, the Minerva approach to selective collection has many advantages for scholars, including collecting all files from a web site, providing better access and the creation of a catalog record. For these reasons, despite its costs, selective collection is recommended to be an important part of the Library's strategy.

Conversely, bulk collection has important scholarly advantages. One is that future generations may value sites that were of no current interest initially. An example might be the personal home page of a student who later becomes prominent. Another is the acquisition of new web sites. With even the most effective selection procedures, with selective collection there will be a delay from when a site is first released until it is recognized. Bulk collection will automatically collect both these categories.

From this analysis, it appears that the Library needs a mixed strategy for selection and collection of web sites.

- Selective selection, for known important sites.

- Bulk selection for selected categories (e.g., .gov sites).

- Bulk collection without selection for other materials.

- Agreements with partners to share the burden and cost of large-scale collecting, e.g., with the National Archives and Records Administration.

4. Use of the Collections for Scholarship and Research

4.1 Analysis of snapshot files by computer

When snapshot files have been downloaded to an archive, they are arranged and stored in a manner designed for use by scholars and for long-term preservation. For access, the organization of these files must recognize the variety of ways in which digital materials might be used for research.

The most basic organization is to store the snapshot files exactly as downloaded. Such raw snapshot files are convenient for analysis by computer. Computers are essential tools for the analysis of huge volumes of information. Programs that read through the archives, searching for specific information or identifying patterns, carry out the first stage of research. Such analyses are best carried out on the original, unmodified source files, rather than the rendered form as seen by a user of the web materials. The files must be organized in a manner that allows rapid processing of large numbers of web sites, based on their logical structure.

As an example of support for this method of access, the Internet Archive has organized its materials for computer analysis. It provides a number of computers for researchers to install and run their own programs.

4.2 Access versions for viewing by scholars

Other scholars will wish to view preserved web sites as they appeared on the date when they were collected. For this, they want to view and interact with the rendered form of the web site, not the underlying source files. The National Library of Australia's Pandora site is a good example of collections that have been organized in this manner. For the Minerva study, a web site was built to provide online access to the sites that had been downloaded. This web site was designed for scholars who wish to experience individual sites as they were originally.

Unless a web site has been designed for mirroring, the snapshot will not render correctly and strange things happen when a user attempts to view it, particularly at a later date. For example, the mirror of the George W. Bush web site defaulted to the Spanish version. To overcome such problems, some of the files have to be edited, to create an access version. Here are some of the editing changes that are needed.

Absolute URLs

A common problem is that any URL that specifies an absolute address will link to the current location with that address, not to the contemporary version in the archive. Internal hyperlinks need to be edited to relative addresses, not absolute addresses.

Executable code

Many files include program code that, for example, displays the current date. Such code needs to be modified to use the date on which the snapshot was taken.

For reasons of cost, it is highly desirable that this editing is carried out entirely automatically by computer program. However, because of the variety and complexity of many sites, inconsistencies will slip through. Yet the costs of manually checking and editing every web page are too high to do routinely. Since the software and hardware systems used to render web sites change with time, specific efforts must be made to render old sites.

4.3 Policies for access

The technical questions of how to organize the collections for use by scholars need to be related to the policies and procedures governing access. This is a complex topic that was not studied during the pilot part of the project.

4.4 Recommendations for the Library of Congress on the use of the collections for scholarship and research

These two types of analysis, by computer and by viewing an access version of the web sites, serve different types of research. The Library will need to provide alternative methods of access.

- Computer analysis of snapshot files.

- Automated editing to create access versions of all selected sites, without manual checking.

- Manual checking and editing of access versions of a few, very important sites.

The first two alternatives are moderately low cost, though the automated editing programs needed for the second option will need continual enhancement. Manual checking, however, is needed to maintain a version of the web site that replicates the original experience. This is very expensive indeed. To be conscientious, every page needs to be viewed, with every combination of options, from a variety of web browsers, on a variety of hardware, with several different operating systems.

Whatever level of editing is carried out, there is every likelihood that, because of technology obsolescence, eventually it will be impossible to retain the original look and feel, and the experience of using the web sites.

5. Information Discovery

5.1 Options for information discovery

Since very large numbers of web sites will be collected and preserved, some form of catalog, index or finding aid is required, but there is no body of experience to know how researchers will use a large archive of web materials.

If catalog records are created they can be item level, one for each web site, or collection level, for groups of sites (such as the group of web sites that cover the 2000 election). The records themselves can be full catalog records or use some shortened form, such as Dublin Core. (The project has not yet studied Dublin Core in this context.) Other means of access that might be provided at much lower cost include searchable lists of URLs, <title> fields extracted from HTML pages, or free text indexes of home pages.

The Minerva pilot provided four forms of information discovery, as described below. All were linked to the access versions of the web sites.

- MARC catalog records.
- List of sites by URL.
- List of sites by title.
- Subject access.

The pilot did not experiment with another possible form of indexing, an automatically generated index of the kind made familiar by the web search engines. This needs serious study as a low-cost and flexible alternative to conventional cataloguing.

5.2 Cataloguing

For the Minerva study, MARC item level catalog records were created for each of the web sites. The software used was OCLC's CORC system, which is web based [8]. Since OCLC's database already contained records for most of the 29 sites that were nominated by the selectors, copy cataloguing was used where available. This went smoothly and the time to create each record was about one hour, comparable to the usual time to create a record for an electronic file.

The completed records were then loaded into the ILS at the Library of Congress. Links have been created from the Minerva web site to the ILS, so that a user of the web site can search in the ILS and then view the archived web sites.

The preliminary study showed that OCLC's CORC software provides a good tool for creating catalog records and that there are no fundamental obstacles to integrating such records into the ILS and the Library's other procedures. However, it illustrated some peculiarities of web sites that would need to be addressed in a production system.

Changes over time

The dynamic nature of web sites poses serious problems. For instance, the Gore and Bush sites changed dramatically when the vice-presidential candidates were chosen, and URLs change frequently for many sites. For such sites, a catalog record, unless it is very generic, will need continual updating or become increasingly inaccurate.

Title

The title of a web site often poses difficulties. The HTML <title> can sometimes be used, but it is often poor. Titles on the rendered version of the home page may change erratically.

Identifiers

Identification of web sites is a continual problem. A given web site may be referenced by several URLs any one of which may change at time.

Navigation

When many archived versions of a web site are available, a user can easily be confused about which version is being displayed. Typically, many versions of the same web site look similar, including the current version accessible over the Internet. Some form of visual identification in desirable, such as framing the archived versions with a suitable banner.

5.3 Recommendations about information discovery

It is difficult to make recommendations about cataloguing strategies because of lack of knowledge of user needs.

- The Library should experiment with various approaches to indexing and cataloguing web sites, including automated indexing, Dublin Core and MARC cataloguing. In particular, the Library needs to study the effectiveness of automated indexing as a low-cost alternative to cataloguing.

- The Library will probably not be able to afford individual catalog records for all web sites that are collected.

- It seems reasonable to provide a browsable list of web sites, but it is not obvious whether to refer to them by URL, subject, a <title> field, or the title from a catalog record where available.

6. Legal Issues

6.1 The legal situation

Despite the anxieties surrounding copyright in electronic materials, the collection of born-digital materials for the benefit of future scholarship is clearly consistent with the Library of Congress's historic duties. These are recognized and supported by the Copyright Act and the legislative history surrounding Section 407 (demand deposit).

For the Library to carry out its responsibility to preserve digital information – most of which is subject to copyright -- the legal framework must be clear and unambiguous. While it is reasonable to assume that most organizations that make information openly available on the web would be willing for the Library of Congress to download copies and keep them for future research, the Library does not currently have the explicit legal right to do so.

In collecting and preserving open access web sites, there appear to be three inter-related questions to address.

- Is it legally correct for the Library of Congress to download web sites and preserve them for the future as an alternative to demanding that publishers deposit materials through the Copyright Office?

- Under what circumstances should the publishers of web sites be informed that their sites are being collected and preserved?

- After web sites have been collected, what policies should determine the scope of access to them?

6.2 Downloading open access materials from the web

Through the system of copyright registration and deposit defined in Chapter 4 of the Copyright Act, the Library of Congress receives a copy of essentially all materials registered for copyright and has the right to demand copies of all materials published in the USA to add to its collections.

With open access materials, the Library is able to acquire materials by downloading, rather than be delivery from the publishers. This is convenient for everybody, but was not anticipated by the Copyright Act. To develop a full program of collecting and preserving open access web sites, in this manner, the Library needs authority for three activities:

- Where materials have been made openly available without restrictions, the Library of Congress will download copies from the web rather than demand copies from the publisher. Moreover, in these cases, the Library will not ask permission before downloading materials for preservation. Because of the volume of materials, this is essential to collecting the materials at a reasonable cost.

In addition to materials that are available without restriction, there are some materials that are available over the Internet, but have restrictions on access. The most common restriction is robot exclusion, where a directory on the web site has a file requesting robots not to access the materials. Examples of sites that use robot exclusion but are important for long-term preservation are some of the national newspapers. These materials should not be downloaded without permission. It is our understanding that the Library could use its rights under the current law to require deposits of such materials.

- The Library of Congress may choose to designate one or more other organizations, at locations other than the Library, to act as its agents to carry out collection and preservation of open access materials on its behalf. As with other digital materials, the Library needs to store copies at remote locations for security. As a practical matter, the Library does not currently have the resources to download and store vast digital collections in a cost-effective manner. Even if it had such resources, the complexity of preserving digital information is so great that the Library needs the flexibility to build collaborative partnerships and share expertise with other leading organizations. The Internet Archive is an example of an organization that might be a possible agent.

- As discussed in Sections 4 and 7, the Library will often make small editorial changes to the materials that it downloads for reasons of access and preservation.

The Copyright Office has offered to work with the Library to make explicit exactly what is required. When this is completed, the Librarian of Congress, with support from the Copyright Office, may need to ask Congress to amend Section 407 of the Copyright Act to permit downloading of open access materials that are on the Internet.

7. Long-term Preservation

7.1 Preservation objectives

It is difficult to guess how future researchers will use preserved collections of web sites, yet strategies for preservation are interwoven with assumptions about why the collections are being preserved. Preservations strategies can be judged against three objectives:

- Preservation of bits. Here the aim is to retain the exact bit sequence of the original.

- Preservation of content. Here the aim is to retain the content (e.g., the words in a text) but not the formats, graphical design, etc.

- Preservation of experience. Here the aim is to preserve the entire experience of interacting with the digital material, including the graphical design and execution of dynamic elements.

The complexity and cost of preservation increase sharply across these three objectives. Preservation of experience is particularly demanding since replacements must be found for hardware and software that is no longer available.

7.2 Preservation strategies

There are several technical strategies for preserving digital information. Each is appropriate for different preservation objectives and methods of access.

Permanent storage

In the past, whether a physical artifact survived depended primarily on the longevity of its materials. Of today's digital media, none can be guaranteed to last for long periods. Some, such as magnetic tape, have a frighteningly short life span before they deteriorate. Others, such as CDs are more stable, but nobody will predict their ultimate life. Ideally, the Library would store its digital materials on permanent storage devices, but unfortunately no permanent media exist with the large capacities needed.

Refreshing

Because all types of physical media on which digital information is stored have short lives, methods of preservation require that the data be copied periodically onto new media. The Library must plan to refresh the collections in this manner. Every few years the data must be moved onto new storage media. From a financial viewpoint this is not a vast challenge. For the next few decades, computing equipment will continue to tumble in price while increasing in capacity. The equipment that will be needed to migrate today's data ten years from now will cost a few percent of the cost today and robots can minimize the labor involved. Refreshing achieves the first objective, preservation of bits.

Migration

Computing formats change continually. File formats of ten years ago may be hard to read. There is no computer in the world that can run programs for computers that were widespread a short time ago. Therefore, in addition to refreshing the raw data, digital preservation must preserve ways to interpret the data, to understand its type, its structure, and its formats. Migration has been standard practice in data processing for decades. Businesses, such as pension funds, maintain records of financial transactions over many years. These records are kept on computers, but the computer systems are changed periodically. Hardware is replaced and software systems are revised. When these changes take place, the data is migrated from computer to computer, and from database to database. The basic principle of migration is that the formats and structure of the data may be changed, but the semantics of the underlying content is preserved. Migration always preserves content and may sometime preserve the experience.

Emulation

Another method that is sometimes suggested is emulation. The dream is to specify in complete detail the computing environment that is required to execute a program. Then, at any time in the future, an emulator can be built that will behave just like the original computing environment. In a few, specialized circumstances this is a sensible suggestion, but in most circumstances, emulation is a chimera. Even simple environments are much too complex to specify exactly. The combination of syntax, semantics, and special rules is beyond comprehension, yet subtle, esoteric aspects of a system are often crucial to correct execution.

7.3 Recommendations about preservation

Several recommendations follow from the previous discussion:

- Keep the original source material and also any altered versions that have been edited for access or by migration.

- Periodically refresh all versions by copying from old media to new.

- On a regular basis, review the formats, protocols, program components, etc, used. If they are becoming obsolete, migrate to the nearest modern equivalent.

- Develop a library of technical specifications, standards and software utilities that describe the current technology used by web sites. This will be a resource for future scholars in puzzling out file types and formats that have become obsolete.

Even with extensive migration, there is no guaranteed way to preserve the experience of using digital materials. For the foreseeable future, preservation of experience will be labor-intensive and require analysis of individual web sites. To analyze individual web sites, digital archeology will be needed, examination of the original source materials by skilled researchers to extract content and experience.

8. The Development of a Production System

8.1 Estimates of volumes

There are no reliable statistics of the number and size of the web sites that the Library of Congress might wish to collect and preserve. Here are some figures that give an idea of the scale.

Number of web sites

For several years, OCLC's Web Characterization Project has gathered annual statistics of the number of web addresses that provide public access. In February 2000, this number was 2,900,000. The growth rate appears to be about 700,000 per year. If the Library of Congress were to select one percent of these sites, the current level would be about 30,000 sites with 7,000 additional per year.

Storage requirements

The average size of sites is equally difficult to predict. The 29 sites that were measured by the present study varied in size from 60 thousand bytes to more that 600 million, with an average of about 60 million bytes. The sizes of the three sites used for detailed study were:

www.whitehouse.gov	618 Mbytes
www.algore2000.com	101 Mbytes
www.georgewbush.com	14 Mbytes

Most of this size comes from files in formats other than text. For instance, the Gore site contains many digitized videos. The text files are relatively small.

With this average size, a single snapshot of 30,000 web sites would require 1,800 gigabytes, or about 2 terabytes. The size of a web site depends heavily on the formats used. Since text in HTML format uses little space, the Internet Archive, which collects all open access HTML pages every month and mounts them online, has a total of fewer than 15 terabytes. If, however, the Library plans to collect other formats, such as video, audio and bit-mapped images, it is clear that the Library will have to store very large quantities of material.

Storage costs

The cost of disk storage is currently ranges from about $25,000 per terabyte for straightforward servers to $100,000 per terabyte for the RAID disks used by the Library of Congress. Although the cost can be expected to fall rapidly, the size of web sites is also likely to grow rapidly. Therefore storage costs will be significant and methods of saving space will be needed. When a web site is collected at frequent intervals, many copies of the same files will often be downloaded. To economize on storage, only one copy should be stored. While greatly reducing the storage, this approach, unfortunately, increases the complexity of the computer systems. The following table summarizes the storage requirements for a single version.

Number of sites collected	30,000
Average size of site	60 Mbytes
Size of 30,000 sites	1.8 terabytes
Storage requirements/year (monthly snapshot)	21.6 terabytes
Storage requirements/year (no duplicates)	5.0 terabytes
Cost per year, per version ($100,000 per terabyte)	$500,000

These figures are all very tentative.

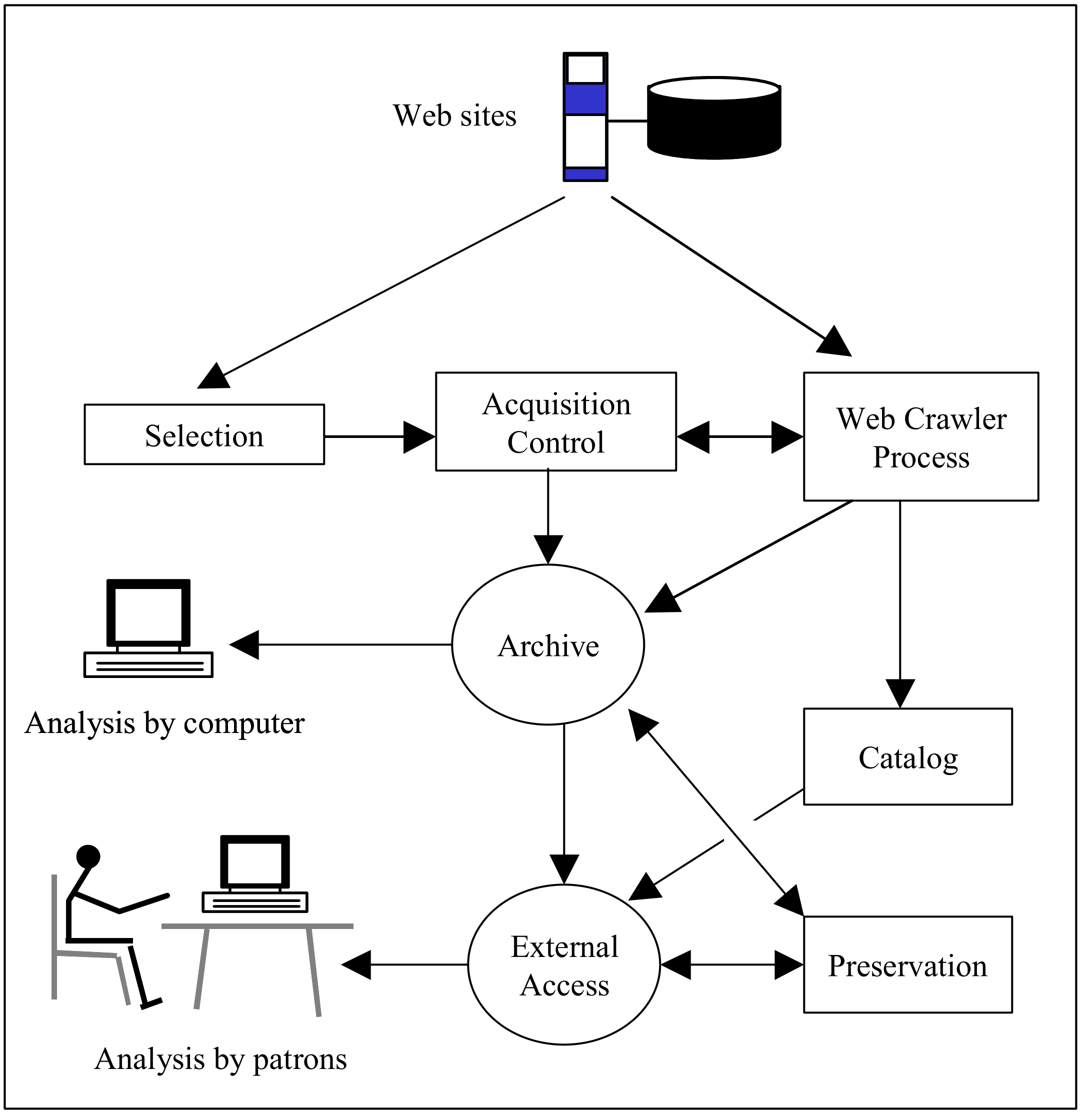
Versions

The previous table shows the cost of storing the snapshot files. Often, however, several versions will be stored. In addition to the snapshot, an access version may be needed. Over time, additional versions will be created for preservation. (See Section 7.)

Storage costs are falling rapidly, as least 30 per cent annually for magnetic disks. This will greatly reduce the costs of storing snapshots collected today and in making new versions for preservation. However, there is every indication that the size of new web sites is growing equally rapidly, so that the budget for storage is unlikely to decline.

8.2 Computer systems

Any system for collecting and preserving web sites will reply heavily on automation. The following figure is an outline of a computer system to collect and preserve web sites at the Library of Congress.



The system divides into several subsystems:

<u>Selection</u>

This provides a librarian with the support tools to examine and select web sites for collection and preservation, and to set parameters such as frequency of collection, the boundaries of the web site, etc.

<u>Acquisition control</u>

This subsystem schedules the collection of snapshots, monitors the performance of the web crawler, and examines logs and error messages. In some aspects it resembles serials check in; other characteristics are like version control in software development.

Web crawler

This mirrors web sites, when instructed by the acquisition control system, and stores them in the archive.

Archive

The archive provides long-term storage for the snapshot files. It also provides access to these files for analysis, usually by computer program.

Catalog system

This is a multi-part system. It includes the tools used to make catalog records (e.g., OCLC's CORC software), interfaces to the ILS catalog, other indexes and finding aids, and interfaces for searching and browsing.

External access

This subsystem edits the snapshot files for use by scholars and researchers. It also builds and maintains the web site(s) used for external access.

Preservation

This subsystem monitors the archive and access versions. It creates new versions of the files for preservation and generates provenance metadata.

Most of the subsystems, but especially the archive, external access and preservation subsystems, need to interface to an identification system to keep track of the many versions of each web site. The Handle system is suitable for this purpose and preliminary work has begun in specifying how it might be used.

8.3 In-house development and out sourcing

The Library has to decide how much of this system to develop and operate in-house and how much to contract out. This is a central system for the scholarly mission of the Library of Congress. Overall responsibility for its design and operation must remain in the Library. Some specialized parts, however, might be subcontracted or carried out with partners. The staffing plans are based on the following assumptions.

- The Library will take overall responsibility for the system architecture and for operating the system.

- Existing software tools will be used wherever possible, including the ILS, CORC and the Handle System.

- The development of the web crawler and its operation will be contracted out to an organization with the specialized expertise that is needed. (Operating the crawler at the Library would require two extra staff.)

- The other systems will use existing software wherever possible, extended and maintained by the Library.

8.4 Staffing for a production system

The principal resources needed for a program of web preservation are staff and equipment.

For this program to succeed, a team of professional staff must be assigned full time to it. To some extent the number of staff assigned will determine the volume of material collected. While a team of 10 is adequate to begin, a larger team would be better. The following estimate is based on an assumption of 30,000 sites being collected with an average frequency of monthly.

Team leader

> The team leader must understand scholarly, technical and librarianship aspects of the program, within the practical realities of the Library of Congress.

Technical.

> As indicated above, a substantial computer system is needed to manage the materials. With the assumptions in Section 7.2, it is estimated that development of this system will require two people for two years, followed by one person to maintain and extend it. In addition, a system administrator, with programming ability, will be needed to operate the system and deal with the numerous exceptions likely to occur. Because the technical shape of web sites changes continually, these tasks will continue indefinitely. (3 technical staff during initial two-year phase; 2 thereafter. The primary technical skill is experience in programming for the web.)

Selection

> Selection has several parts. One is interaction with the curatorial divisions to identify web sites that are candidates for preservation. In addition, there is a need for continual observation to identify opportunities for preservation that fall beyond the scope of the individual curatorial departments. Once sites have been identified as candidates for preservation, it becomes necessary to analyze them to decide on frequency of collection and to look for any peculiarities. Conversely, it is also necessary to monitor the sites that are being collected to decide when to stop collecting or change the collection parameters. (3 librarians with collaboration from the curatorial divisions.)

Cataloguing and indexing

> The level of staff needed for cataloguing and indexing depends on the level of cataloguing and whether individual sites are catalogued or collections of sites. Assuming that each librarian catalogs 750 sites per year, to catalog 7,000 new sites and 7,000 changed sites per year would require nearly 20 staff. The recommendation

is that fewer staff be assigned to this program and that most sites be catalogued at a collection level. (2 librarians with experience of cataloguing digital materials.)

Preservation and access to the materials

Organizing the materials for access, including maintenance of the web site(s) is a substantial and ongoing task that is difficult to estimate. It is closely related to preservation. Preservation of digital materials is a major task for the Library of which this is just a small part. (2 staff, in addition to the staff involved with other preservation activities at the Library.)

Behind this staffing plan is an assumption that the Library will be steadily developing staff expertise in managing digital materials as recommended in the recent study by the National Research Council [1]. Web preservation can make use of and contribute to the broader program in areas such as training and the development of technical skills.

References

1. National Research Council, *A Digital Strategy for the Library of Congress*. National Academy Press, 2000. http://www.nap.edu/books/0309071445/html/.

2. National Research Council, *The Digital Dilemma, Intellectual Property in the Information Age*. National Academy Press, 2000. http://www.nap.edu/html/digital_dilemma/.

3. OCLC Office of Research, *Web Characterization Project*. http://wcp.oclc.org/main.htm

4. National Library of Australia, *Pandora Archive – Preserving and Accessing Networked Documentary Resources of Australia*, 1996-. http://pandora.nla.gov.au/pandora/.

5. The Internet Archive, *Building an Internet Library*, 1996-. http://www.archive.org/.

6. Allan Arvidson, Krister Persson and Johan Mannerheim, The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of web pages. *66th IFLA Council and General Conference*, Jerusalem, Israel, 13-18 August 2000. http://www.ifla.org/IV/ifla66/papers/154-157e.htm.

7. *HTTrack, the Web Mirror Utility*. http://httrack.free.fr/.

8. OCLC, *Cooperative Online Resource Catalog*. http://www.oclc.org/oclc/corc/

Appendix A – Web sites analyzed

NAME	URL	TIME	LINKS	FILES	BYTES	ERRORS
Al Gore For President	http://www.gore2000.com	15M, 58S	1,451	1,449	101,316,632	1E
Alan Keys for President	http://www.keyes2000.com	2M, 57S	700	699	10,922,427	21E
American Univ. Campaign Finance	http://www1.soc.american.edu/campfin/index.cfm	2M, 21S	227	226	46,233,840	31E, 2M Robots.txt
Asian-Pacific Economic Cooperation	http://www.apecsec.org.sg/	3H, 48M, 8S	4,055	4,051	239,622,409	135E
Bill Bradley for President	http://www.billbradley.com	4S	20	19	191,866	None
Census Bureau Economic Reports	http://www.census.gov/epcd/www/econ97.html	17M, 32S	4,109	4,108	94,120,855	24E, 2M Robots.txt
China Law	http://www.qis.net/chinalaw/	12S	51	47	110,823	1E
Countdown to the Millennium	http://www.countdown2001.com/index.htm	6M, 4S	1,217	1,080	11,760,535	110E
Daily Nation	http://www.nationaudio.com/News/Daily/Nation/Today/	7M, 26S	1,883	1,124	8,854,844	28E, 755W
E Law - Murdoch Univ. Elec. Law Journal	http://www.murdoch.edu.au/elaw/	11M, 39S	1,515	1,494	46,585,632	21E, 2M Robots.txt
Earth Defense	http://www.edf.org/	40M, 17S	3,906	3,583	95,847,483	83E, 571W
Economic Policy Institute	http://www.epinet.org/epihome.html	49M, 29S	7,276	7,274	62,187,206	27E
George W Bush For President	http://www.georgewbush.com	16M, 46S	292	261	13,975,307	18E, 37W, 313M
Hillary 2000 for New York	http://www.hillary2000.org/	5M, 54S	1,302	1,301	31,268,940	36E

NAME	URL	TIME	LINKS	FILES	BYTES	ERRORS
History and Politics Out Loud	http://www.hpol.org	7S	17	16	59,994	None
Inter-Parliamentary Union	http://www.ipu.org	18M, 15S	1,651	1,650	28,703,312	1E, 2M Robots.txt
Internet Scam Busters	http://www.scambusters.org/index.html	38S	182	181	2,144,641	2E Robots.txt
John McCain for President	http://www.mccain2000.com	2M, 39S	269	268	2,605,516	43E
Julian Samoa Research Institute	http://www.jsri.msu.edu/	16M, 10S	2,007	1,996	84,789,507	30E
National UFO Reporting Center	http://www.nwlink.com/~ufocntr/	12S	5	4	76,025	None
National Women's History Project	http://www.nwhp.org	51S	297	287	1,650,896	9E
Panda in Zoos	http://www.chinaunique.com/Panda/pandazoo.htm	37S	52	51	866,191	None
Pat Buchanan for President	http://www.gopatgo2000.com	1H, 12M, 39S	1,676	1,661	18,892,263	12E
Sierra Club	http://sierraclub.org	1H, 40M, 13S	14,213	13,719	268,908,949	477E, 27W, 2M Robots.txt
Souleyes Magazine	http://www.souleyes.com/	1M, 46S	552	551	10,570,105	None
Virginia Native Plant Society	http://www.wnps.org	58S	257	256	4,657,088	5E
White House	http://www.whitehouse.gov	2H, 32M, 59S	23,604	23,601	617,512,801	1,382E, 59W
World Development Sources	http://www.wds.worldbank.org	36S	162	138	14,266,604	8E, 26M
Year 2000	http://www.year2000.com	7M,35S	1,829	1,825	27,739,981	40E

American Customer Satisfaction Index	http://www.bus.umich.edu/research/ngr/acsi.html	URL not found
China Today-Law	http://www.chintoday.com/law/a.htm	URL not found
Global Legal Information Network	http://lcweb2.loc.gov/law/GLINv1/GLIN.html	Did not create mirror
International Constitutional Law	http://www.uni-wuerzburg.de/law/index.html	Timed out
Native American Day	http://native-americans.com	URL not found
Rose Bowl Parade 2000	http://www.caribbeaninformation.com/football/football.html#Rose	URL not found